DOCUMENT RESUME

ED 385 562 TM 023 984

AUTHOR	Sebrechts, Marc M.; And Others
TITLE	Machine-Scorable Complex Constructed-Response Quantitative Items: Agreement between Expert System and Human Raters' Scores. GRE Board Professional Report No. 88-07aP.
INSTITUTION	Educational Testing Service, Princeton, N.J.
SPONS AGENCY	Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-91-11
PUB DATE	Apr 91
NOTE	57p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Algebra; *Automation; *College Students; *Computer Assisted Testing; *Constructed Response; Evaluators; *Expert Systems; Higher Education; Mathematics Tests; Scores; *Scoring; Testing Programs; Test Scoring Machines; Word Problems (Mathematics)
IDENTIFIERS	Experts; Graduate Record Examinations; Large Scale Programs

ABSTRACT

This study evaluated agreement between expert system and human scores on 12 algebra word problems taken by Graduate Record Examinations (GRE) General Test examinees from a general sample of 285 and a study sample of 30. Problems were drawn from three content classes (rate x time, work, and interest) and presented in four constructed-response formats (open-ended, goal specification, equation setup, and faulty solution). Agreement was evaluated for each item separately by comparing the system's scores to the mean scores taken across five content experts. Results showed the expert system to produce scores for all responses and to duplicate the judgments of raters with reasonable accuracy; the median of the 12 correlations between the system and human scores was .88, and the largest average discrepancy was 1.2 on a 16-point scale. No obvious differences in scoring agreement between constructed-response formats or content classes emerged. Ideas are discussed for further research and development concerning the use of expert scoring systems in large-scale assessment programs and in interactive diagnostic assessment. Seven tables and 2 figures present study data. Three appendixes present item stems, transcription rules and examples, and the scoring rubric and keys. (Contains 22 references.)
(Author/SLD)

GRE®

RESEARCH

Machine-Scorable Complex Constructed-Response Quantitative Items: Agreement Between Expert System and Human Raters' Scores

Marc M. Sebrechts
Randy Elliot Bennett
and
Donald A. Rock

April 1991

GRE Board Professional Report No. 88-07aP
ETS Research Report 91-11

Educational Testing Service, Princeton, New Jersey

Machine-Scorable Complex Constructed-Response Quantitative Items: Agreement Between Expert System and Human Raters' Scores

Marc M. Sebrechts
Randy Elliot Bennett
and
Donald A. Rock

GRE Board Report No. 88-07aP

April 1991

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

Educational Testing Service, Princeton, N.J. 08541

The Graduate Record Examinations Board and Educational Testing Service are dedicated to the principle of equal opportunity, and their programs, services, and employment policies are guided by that principle.

Acknowledgments

This project was made possible by the concerted efforts of many individuals working in and out of Educational Testing Service (ETS). Subject recruitment and data collection were organized by Hazel Klein and Terri Stirling and implemented by the staff of the ETS Field Services Offices. Steve Anacker, Beth Brownstein, Pat Kenney, and Kathy Martin--all ETS mathematics test development staff members--consulted on developing the scoring rubric and keys and provided an initial corpus of correct and incorrect problem solutions. These individuals, with the addition of Dan Richman, also scored examinee solutions. Jim Spohrer and Lael Schooler played central roles in helping ETS to adapt GIDE to the algebra domain, and Doug Frye offered useful comparisons with a similar study involving MicroPROUST. Students at the Catholic University of America furnished pilot data and served as transcribers and data managers. The Graduate Record Examinations (GRE) Board supplied the financial support to carry out the project, and three ETS reviewers offered helpful comments on earlier drafts of this report. Finally, appreciation is expressed to Henry Braun for his early recognition of the potential value of this technology and for his continued support.

Abstract

This study evaluated agreement between expert system and human scores on 12 algebra word problems taken by GRE General Test examinees. Problems were drawn from three content classes (rate x time, work, and interest) and presented in four constructed-response formats (open-ended, goal specification, equation setup, and faulty solution). Agreement was evaluated for each item separately by comparing the system's scores to the mean scores taken across five content experts. Results showed the expert system to produce scores for all responses and to duplicate the judgments of raters with reasonable accuracy; the median of 12 correlations between the system and human scores was .88 and the largest average discrepancy was 1.2 on a 16-point scale. No obvious differences in scoring agreement between constructed-response formats or content classes emerged. Ideas are discussed for further research and development concerning the use of expert scoring systems in large-scale assessment programs and in interactive diagnostic assessment.

Machine-Scorable Complex Constructed-Response Quantitative Items: Agreement Between Expert System and Human Raters' Scores

Constructed-response items, particularly those that call for an extended or "complex" response (Bennett, in press), are often argued to be more effective than multiple-choice questions for educational assessment (Fiske, 1990; Frederiksen & Collins, 1989; Guthrie, 1984; Nickerson, 1989). The essence of this argument is that constructed-response questions more faithfully replicate the tasks examinees face in academic and work settings. This increased fidelity is said to engender better measurement of higher-order skills, permit responses to be evaluated diagnostically according to both the processes used to arrive at a solution and the degree of correctness, and communicate to teachers and students the importance of practicing these "real-world" tasks.

The potential benefits of complex constructed-response questions can be realized in large-scale testing programs to limited degrees and, often, only at the substantial costs associated with employing human readers. For example, the College Board's Advanced Placement Program annually gathers, trains, and houses hundreds of secondary school teachers and college professors to score hundreds of thousands of essays, designs for laboratory experiments, calculus solutions, and computer programs (College Entrance Examination Board, 1988).

Though a short time ago it might have been implausible for responses of such complexity and variety to be graded by computer, recent work with expert systems--programs that emulate the behavior of a human master--suggests considerable progress. In one study using open-ended items from the Advanced Placement Computer Science examination, an expert system was able to score a significant portion of the solutions presented and to generally duplicate human judges' partial credit scores and diagnostic analyses (Bennett, Gong, Kershaw, Rock, Soloway, & Macalalad, 1990). In follow-up studies using a more constrained constructed-response format, the expert system scored the overwhelming majority of student responses. Its scores agreed highly with a human rater and measured essentially the same underlying attribute as the standard examination (Bennett, Rock, Braun, Frye, Spohrer, & Soloway, 1990; Braun, Bennett, Frye, & Soloway, 1990). These advances make plausible a range of possibilities, from programs that work in tandem with human graders in scoring large volumes of student productions to interactive assessment systems that diagnostically analyze examinees' responses (Bennett, in press).

Several questions can be raised about the analyses produced by such systems, including the accuracy of scores, relations with other ability indicators (particularly multiple-choice tests), and the extent to which scores and qualitative diagnoses can be aggregated across questions to produce diagnostic inferences more general than those based on only a single item. This paper

addresses one of these issues--expert system scoring accuracy--by examining agreement with human judges on constructed-response quantitative items adapted from the GRE General Test. Reasonably high agreement with human graders is essential if such systems are to be used credibly in conjunction with, or in place of, the judgments of content experts.

Method

Subjects

Subjects were participants in a series of studies concerned with automated scoring of constructed-response algebra word problems. The main sample was drawn from a pool of over 50,000 examinees taking a single form of the GRE General Test administered nationally in June 1989. Examinees living within approximately 30 miles' driving distance of an ETS field office were identified and asked by letter to participate for a fee in research to develop instructionally relevant test items.[1] Expressions of interest were received from 1,236 of 3,244 individuals contacted. Respondents were removed from consideration if they were not on the original mailing list, if their citizenship could not be determined from their GRE registration data, or if they no longer lived in the regions served by the ETS offices. From the remaining group, up to the first 100 persons from each region were selected, with some individuals replaced to produce within-region samples composed of citizens and non-citizens in proportions similar to the General Test population. Attempts were made to schedule the resulting 684 people for testing. Of those individuals, 285 participated. From this set, a subsample of 30 examinees was randomly selected for the current investigation.

Table 1 presents background data on the samples and the population taking the June General Test. For each variable, the main sample and June administration values were compared via a two-tailed z-test with alpha set at .05, treating the June value as a population parameter. As can be seen, the main sample differed somewhat from the population. The sample's General Test performance was significantly, though not dramatically, higher (by .4, .3, and .3 standard deviations, for verbal, quantitative, and analytical, respectively), and the most notable of several statistically significant demographic differences was in a greater proportion of non-Whites. The study sample was a random extraction from this main sample and, consequently, compares to the population in similar ways.

Table 1

Background Data for Study Samples

Variable	June 1989 Population	Main Sample	Study Sample
N	50,548	285	30
General Test Performance			
Verbal Mean(SD)	476(122)	527(132)*	544(113)
Quantitative Mean (SD)	532(140)	573(141)*	562(131)
Analytical Mean (SD)	513(132)	558(129)*	546(130)
Percentage Female	55%	60%	60%
Percentage Non-White[a]	16%	28%*	23%
Percentage U.S. citizens	79%	85%*	83%
Undergraduate Major			
Business	4%	2%	7%
Education	14%	6%*	14%
Engineering	13%	12%	7%
Humanities/Arts	14%	21%*	21%
Life Sciences	18%	18%	18%
Physical Sciences	10%	9%	7%
Social Sciences	18%	24%*	21%
Other	9%	10%	4%
Intended Graduate Major			
Business	2%	2%	4%
Education	18%	12%*	11%
Engineering	10%	9%	7%
Humanities/Arts	8%	9%	11%
Life Sciences	16%	15%	22%
Physical Sciences	8%	9%	7%
Social Sciences	13%	19%*	15%
Other	11%	9%	11%
Undecided	15%	18%	11%

* $p < .01$, two-tailed z-test of main sample value with total test population parameter.

[a]U.S. citizens only.

Instruments

Constructed-response items. Items were adapted from standard, five-option multiple-choice algebra word problems taken from disclosed forms of the General Test quantitative section administered between 1980 and 1988. An initial pool of 20 items was selected from six algebra problem classes: rate x time, work, interest, graduated rate, percent, and probability. These classes were chosen because they appeared similar in the concepts and procedures used in problem solving and, thus, might make for greater efficiency in developing the knowledge bases required for analyzing students' responses. In addition, the classes tended to contain problems whose solutions could be broken into components amenable to partial credit scoring and diagnostic analysis.

The knowledge bases were developed--and the final item set selected--through the following steps. First, four ETS mathematics test developers were asked to specify as many correct and incorrect solutions as possible to open-ended versions of the problems. Next, 50 undergraduate students attending the Catholic University of America (CUA) solved the word problems and also equivalent equation items, providing a basis for separating procedural from conceptual problem-solving errors. Finally, 10 CUA undergraduates were asked to work subsets of the questions aloud so their problem-solving approaches could be more easily identified. From these three data sets the procedural and computational knowledge specific to each algebra word problem class, the knowledge common across classes, and the general and specific conceptual errors students typically made in responding were identified.

This information base was used to select three items--one from each of the rate x time, interest, and work classes--that were solved by similar sets of equations, were of intermediate difficulty (in their multiple-choice forms) to permit frequent error diagnosis, and were typically solved by equation as opposed to verbal approaches. Next, three isomorphs were written for each prototype by an ETS test developer, producing 12 items (see Appendix A). Isomorphs were intended to differ from the prototype in surface characteristics only--for example, in topic (filling a tank vs. sending characters to a printer, determining percent profit instead of simple interest), and linguistic form, but not in underlying conceptual structure.

For each problem set, the four isomorphs (i.e., the prototype and its three surface variants) were cast into one of four formats, such that each isomorph appeared in a different format. The formats differed in the degree of constraint placed on the response, one dimension along which item formats can be readily differentiated (Bennett, Ward, Rock, & LaHart, 1990). This variation permitted the accuracy of expert system analyses to be explored as a function of response constraint. Second, it held open the possibility, to be addressed in other work, that

the formats might be used in concert to identify the level of constraint needed by a particular examinee in solving a class of problems.

The first format, open-ended, presented the examinee with only the problem stem. This widely used arrangement was selected because its lack of constraint brings it closest to "real-world" problem solving. This loose structure also makes responses relatively difficult to evaluate, permitting the limits of expert system accuracy to be explored. In goal specification, the second format, the problem stem, a list of givens, and a list of unknowns, or goals, are provided. Still more structure is offered by equation setup, which gives the unknowns and the equations needed to derive them. Both goal specification and equation setup were created for this study, predicated on "intention-based diagnosis" (Johnson & Soloway, 1985; Johnson, 1986), one theoretical approach to automated analysis. The fourth format, faulty solution, presents the problem stem and an incorrect solution for the student to correct. This arrangement was based on research in the computer science domain, which found the item type to be machine scorable and to measure the same trait as open-ended questions (Bennett, Rock, Braun, Frye, Spohrer, & Soloway, 1990; Braun et al., 1990). Figure 1 gives examples of the formats for a set of isomorphs.

Once the final item selection was completed, the general and specific knowledge relevant to these items was abstracted from that developed for the initial 20-item set. This knowledge was next encoded in machine-usable form as three separate knowledge bases (one for each problem set). The three knowledge bases were progressively refined using additional pilot data from CUA undergraduates, made-up solutions, and responses from 12 examinees randomly chosen from the sample of 255 subjects (285-30) not used for this study.

Expert system. The expert system was GIDE, a batch processing laboratory tool designed in earlier versions to detect student errors in statistics and automotive mechanics problems (Sebrechts, LaClaire, Schooler, & Soloway 1986; Sebrechts, Schooler, LaClaire, & Soloway, 1987; Sebrechts & Schooler, 1987). GIDE follows the theory of intention-based diagnosis (Johnson & Soloway, 1985; Johnson, 1986), which was developed to diagnose students' bugs in Pascal programs. Intention-based diagnosis attempts to identify a student's aims in solving a problem and to interpret errors as failures to carry out aspects of the intended solution. As such, the approach tries to explain mistakes in terms of the student's conceptual framework.

Figure 1
Work Isomorphs in Four Item Formats

Open-Ended

How many minutes will it take to fill a 2,000-cubic-centimeter tank if water flows in at the rate of 20 cubic centimeters per minute and is pumped out at the rate of 4 cubic centimeters per minute?

ANSWER:_____

Goal Specification

One of two outlets of a small business is losing $500 per month while the other is making a profit of $1750 per month. In how many months will the net profit of the small business be $35,000?

Givens
Profit from Outlet 1 = _____
Profit from Outlet 2 = _____
Target Net Profit = _____

Unknown
Net Monthly Profit = _____
Months to Reach Target Net Profit = _____

ANSWER:_____

Equation Setup

A specialty chemical company has patented a chemical process that involves 2 reactions. Reaction 1 generates 24 grams of molecule B per minute and reaction 2 consumes 5 grams of molecule B per minute. If 4,560 grams of molecule B are desired as a product of this process, how many minutes must it continue?

Equations that Will Provide a Solution:

Net Amount of B Per Minute = Amt. Produced by Reaction 1 + Amt. Produced by Reaction 2
Time for Desired Amount of B = Desired Amount of B/Net Amount of B Per Minute

Your Solution:

ANSWER:_____

Faulty Solution

$3.50 in tolls is received each minute at an automated toll booth while the rate at a booth with an operator is $2.80 each minute. How many minutes elapse before the automated booth receives $14.00 more in tolls than does the person-operated booth?

Tolls per Minute = $3.50/min + $2.80/min
Tolls per Minute = $6.30/min
Time for $14 lead = $14/$6.30 per minute
Time for $14 lead = 2.22 minutes

Your Corrected Solution:

ANSWER:_____

Note. Print size is reduced and page arrangement modified for publication purposes.

GIDE's approach to intention-based diagnosis is implemented by means of goal-plan analysis. For each problem, GIDE has a specification that identifies both the "given" information and the goals into which the problem has been decomposed, where a goal is one of several objectives to be achieved in reaching a solution (e.g., an intermediate result).[2] To be considered correct, a solution must satisfy each goal. GIDE attempts to discover how the student solution satisfies a particular goal by testing it against a series of alternative correct plans (i.e., stereotypical procedures) drawn from its knowledge base. (See Figure 2 for examples of goals and plans.) If no matching plan is found, GIDE attempts to discover the nature of the discrepancy by testing plans that incorporate conceptual errors commonly made in achieving that goal or bug rules that represent more general mistakes. When no plan, buggy or correct, can be matched, the goal is considered missing.

Though similar to the original intention-based strategy used in programming, the approach employed in the current adaptation of GIDE has important differences. In the case of programming, a student solution must be exact and exhaustive. All steps must be present and written according to a highly constrained syntactic form. Insofar as a set of symbols--without reference to computational value--must be present for a student's program to be considered correct, pieces of program code can be treated as relatively independent entities.

These constraints do not normally hold for algebra word problems. Although solutions must reflect the satisfaction of a sequence of goals, there is substantial variability in the forms the solutions can take. Students frequently leave out intermediate components and include extraneous steps. In addition, solution components are rarely independent: the result of a particular computation can influence all subsequent steps. Finally, the student's attempt to satisfy a particular goal can include numerical values that deviate in many ways from the expected result.

GIDE's inference mechanism has been modified to cope with this variability. First, GIDE includes means for recognizing intermediate steps that are not explicitly represented. For example, if the computation to produce a travel distance is not present but that distance correctly appears in a subsequent rate calculation, GIDE infers--barring evidence to the contrary--that the missing step was correctly performed. Second, GIDE will carry through erroneous values. In this way, it can determine if the solution is correctly structured given a particular computational error. Lastly, GIDE uses contextual information to determine the source of erroneous values. When a clock time is expected, GIDE searches for AM/PM confusions. If the form of an equation matches the current plan, but the result does not, GIDE infers a computational error.

Figure 2

A Problem Decomposition Showing a List of Goals for Solving the Problem and Two Correct Plans and One Incorrect Plan for Achieving the First Goal

On a 600-hundred mile motor trip, Bill averaged 45 miles per hour for the first 285 miles and 50 miles per hour for the remainder of the trip. If he started at 7:00 a.m., at what time did he finish the trip (to the nearest minute)?

Goals

1. Find the time for the first part of the trip.

   Correct Plan #1: time1 = distance1/rate1

   Correct Plan #2: distance1/X hours = rate1/1 hour
   distance1 * 1 hour = rate1 * X hours
   distance1/rate1 = X hours.

   Buggy Plan #1: time1 = distance1 * rate2

2. Find the missing distance for the second part.

3. Find the time for the second part.

4. Add the times for the two parts to get a total time.

5. Add the total trip time to the starting time.

After processing each student solution, GIDE's analytical mechanisms are used to produce a brief bug report and a partial credit score. The bug report identifies the nature and location of errors detected. Because of its experimental nature, GIDE's bug reports are relatively unrefined, giving only enough detail to permit verification of an error's existence by an independent source. In any operational implementation, these descriptions would need to be carefully crafted to communicate clearly the nature of the error and perhaps a method for resolving it.

Procedure

Items were presented to the total examinee sample (n=285) in individual and small group sessions conducted at ETS field offices. Examinees were asked to complete the problems at their own pace, though a one-hour period was suggested.

Because astute examinees might recognize the presence of isomorphs and transfer solution processes from problems in one format to their isomorphs in another (Reed, 1987; Reed, Dempster, & Ettinger, 1985), several steps were taken. First, the three problems of a format were presented together and examinees were asked to complete questions in sequence without referring to their earlier work. Second, to reduce recall, each format was separated by two General Test multiple-choice items taken from quantitative content areas other than interest, rate x time, and work. Finally, items were presented in two orders given to random halves of the sample at each location. The orders were (1) most constrained to least constrained (i.e., faulty solution, equation setup, goal specification, open-ended) and (2) the reverse. These orders permitted some degree of control over an order effect in which solutions to the more constrained items might provide guidance in solving the less constrained ones.

Since one of the major benefits of constructed response is its proximity to real-world problem solving, minimal constraints were imposed on the form of students' problem solutions. Examinees were asked to write legibly, place the steps needed to solve each problem in sequence on a lined answer sheet, and compute all results to two decimal places. In addition, it was suggested that only one equation be placed on a line and that the units associated with each quantity be included.

Items were presented in paper-and-pencil format as an efficient means of data collection.[3] Handwritten responses were then converted to machine-readable form according to transcription rules (see Appendix B for the rules and examples of original and transcribed responses). These rules were constructed to place the student's response into a format amenable to machine analysis without changing its substance. These format changes primarily involved arranging solution elements in a linear sequence, translating each line to a syntactically correct equation (e.g., allowing only one "equals" sign per line), and ignoring illegible portions. After

transcription, all solutions were checked for rule violations by a second transcriber. The reproducibility of the transcription process was tested by having two coders independently transcribe the same random sample of 14 examinees' responses to each of the 12 problems, making GIDE score both sets of responses, and computing the product-moment correlation between the two sets. This analysis produced a median correlation of .96, with the lowest value at .87 and the 11 remaining ones above .90.

A scoring rubric and set of keys were developed in consultation with test development content experts (see Appendix C). The rubric was derived from the goal-plan analysis to give scores a principled, cognitive basis. Full credit was awarded if all goals were achieved, suggesting the student was able to decompose the problem, correctly structure each goal, and compute its solution. Credit was deducted differentially depending on the errors detected for each goal. The largest deduction was made for missing goals, because these absences suggest the student was unaware that addressing the goal was necessary to achieve a correct result. Less credit was deducted for structural bugs, because such errors suggest both recognition of the goal's importance and a coherent, though incorrect, attempt to solve the goal. The smallest deduction was for computational errors, which may imply failures in basic calculation skills or procedural "slips" (Matz, 1982). Score scales for the items were based on the number of goals required for solution. Isomorphs developed from the work prototype contained two goals and were scored on a 0-6 scale. Problems based on the interest item were decomposed into three goals and scored on a 0-9 continuum. A 0-15 scale was employed for the rate items, which required solving five goals for a correct response.

GIDE implements the scoring rubric by determining which errors, from a list of 33 general and 19 problem-specific bugs, exist in the solution. It then subtracts the appropriate points depending upon whether the error was structural, computational, or indicative of a missing goal. GIDE's implementation was tested and refined repeatedly using real and made-up solutions from sources other than those used for the agreement analysis.

Data Analysis

GIDE's scoring accuracy was assessed by determining its agreement with human content experts. Graders were five ETS test developers, three of whom held a Ph.D., one a master's, and one a bachelor's degree, in mathematics or mathematics education.[4] The median number of years mathematics teaching experience was 5 (ranging from 2 to 17), and the median years in test development was 1.5 (ranging from two months to 6 years).

The human raters' scoring was conducted as follows. Raters were given the finalized rubric to review several days in advance. At the session, an experimenter explained the rubric,

reviewed the keys for the first set of isomorphs, and demonstrated scoring for several examples. The procedure included indicating a total item score, a score for each goal, and the nature and location of any error(s), with the last piece of information to assist in subsequently analyzing the causes of score disagreements. Raters next practiced scoring using examinee solutions not employed in the agreement analysis. Finally, operational grading began with each rater being given a randomly ordered set of the original student responses to the first isomorph. When all raters had scored all four isomorphs, the training process was repeated for the next set, the set was graded, and so on until all three item sets had been evaluated.

GIDE's accuracy was assessed by comparing its score for an examinee with the corresponding mean score taken across raters. This mean score is conceptually similar to classical test theory's "true" score, the mean of many independent observations of the same performance and, as such, is an approximation of what the "correct" item score for an examinee should be. To evaluate the reliability of this criterion, the models and methods of generalizability theory were employed (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). A three-way repeated measures analysis of variance (with the between-group effect only for persons) was used to estimate the variance components of the following mixed model:

$$Y_{ijk} = \mu + R_j + F_k + \pi_i + RF_{jk} + \pi R_{ij} + \pi F_{ik} + \pi RF_{ijk}$$

where Y_{ijk} is the score assigned to the ith person by the jth rater for the kth format, π is the person effect, R, the rater effect, a random facet presumed to be sampled from an infinite universe of raters, and F, format, a fixed effect representing a universe of four item formats. The data analyzed were the scores assigned by each of the five raters to 30 examinees' responses to each of the four isomorphs (formats) within a problem set. Because the three problem sets were graded on different scales, this analysis was performed separately for each set.

A generalizability coefficient for each problem set was generated as per Thorndike (1982, p. 165-167):

$$r_{xx} = \frac{\sigma_\pi^2 + \frac{\sigma_{\pi F}^2}{k}}{\sigma_\pi^2 + \frac{\sigma_{\pi F}^2}{k} + \frac{\sigma_{\pi R}^2}{j} + \frac{\sigma_{\pi RF}^2}{j*k}}$$

where σ_{π}^2 is the variance component for persons, $\sigma_{\pi F}^2$ the component for the person-by-format interaction, $\sigma_{\pi R}^2$ the component for person-by-rater interaction, $\sigma_{\pi RF}^2$ the component for the person-by-rater-by-format interaction, and $\underline{k}$ and $\underline{j}$ the number of formats and raters, respectively. The resulting coefficient indicates the expected correlation between the mean scores obtained here and those that would be obtained if the experiment were rerun with the same four formats (items) and five new randomly sampled raters each grading all formats.

Agreement between GIDE and the raters was computed in several ways. First, for each item, the Pearson product-moment correlation between GIDE's scores and the mean scores taken across raters was calculated. Next GIDE's mean score for each item was compared with the item mean for each rater to identify any systematic leniency or strictness. Third, the discrepancies between GIDE's scores and the rater means for each examinee were evaluated, first for all responses and then within perfect and imperfect solutions separately.

Results

Results of the variance components analysis for rater agreement are presented in Table 2. Generalizability coefficients for the mean ratings taken across judges and formats were .98, .97, and .99 for the two-goal, three-goal, and five-goal problem sets, respectively. The largest variance component was consistently the subject-by-format interaction, an indication of differences in the patterns of examinee scores across formats.[5] Subject differences also constituted a large component, though for the three-goal problems this component was exceeded by the format effect. Given the absence of any similar format effect in the other problem sets, this effect may be due more to a lack of parallelism in one of the isomorphs (which were nested within formats) than to differences among the formats themselves. Finally, in all three cases, the rater-by-format interaction was trivial, suggesting that the ordering of rater means remained relatively constant regardless of format.

Shown in Table 3 are the grand mean scores taken across raters and examinees for each item, the mean of the score standard deviations (where each standard deviation was taken across examinees for a rater), and the number of perfect responses (i.e., where the grand mean was equal to the score maximum). It is immediately evident that the items were relatively easy overall. For the two-goal problems and three of the four three-goal problems, the majority of responses were perfect. Scores for the five-goal problems were distributed somewhat more evenly, though still clustered in the top third of the score scale.

Table 2

Variance Components for the Scores of Five Human Raters (N =30)

Two-Goal Problems (scale = 0-6)						
Variance Component	Sum of Squares	df	Mean Square	F	P	Variance Estimate
Subject	471.55	29	16.26	70.40	<.01	.80
Rater	1.11	4	.28	1.20	NS	.00
Format	31.85	3	10.62			.04
Subject-by-rater	26.79	116	.23			.06
Subject-by-format	453.81	87	5.22	20.50	<.01	.99
Rater-by-format	3.55	12	.30	1.16	NS	.00
Subject-by-rater-by-format	88.55	348	.25			.25

Three-Goal Problems (scale = 0-9)						
Variance Component	Sum of Squares	df	Mean Square	F	P	Variance Estimate
Subject	565.88	29	19.51	39.64	<.01	.95
Rater	9.90	4	2.48	5.03	<.01	.02
Format	947.81	3	315.94			2.00
Subject-by-rater	57.10	116	.49			.12
Subject-by-format	1248.15	87	14.35	23.09	<.01	2.75
Rater-by-format	19.55	12	1.63	2.62	<.01	.03
Subject-by-rater-by-format	216.25	348	.62			.62

Five-Goal Problems (scale = 0-15)						
Variance Component	Sum of Squares	df	Mean Square	F	P	Variance Estimate
Subject	3181.37	29	109.70	214.74	<.01	5.46
Rater	2.94	4	.74	1.44	NS	.00
Format	66.16	3	22.05			.00
Subject-by-rater	59.26	116	.51			.13
Subject-by-format	2490.04	87	28.62	70.74	<.01	5.64
Rater-by-format	11.01	12	.92	2.27	<.01	.02
Subject-by-rater-by-format	140.79	348	.40			.40

Table 3

Summary Statistics for Scores Awarded by Five Raters (N = 30)

Two-Goal Problems				
Problem	Format	Grand Mean	Mean SD	Number of Perfect Responses
#1 2000 cc tank	Open-ended	5.1	1.9	22
#2 Small business	Goal specification	5.7	.6	21
#3 Chemical company	Equation setup	5.5	1.3	21
#4 $3.50 in tolls	Faulty solution	5.5	1.1	23

Three-Goal Problems				
Problem	Format	Grand Mean	Mean SD	Number of Perfect Responses
#1 Investment fund	Open-ended	8.3	2.0	24
#2 Load of cement	Goal specification	8.5	1.2	23
#3 Graphics designer	Equation setup	8.8	.6	26
#4 Active ingredient	Faulty solution	5.7	2.9	9

Five-Goal Problems				
Problem	Format	Grand Mean	Mean SD	Number of Perfect Responses
#1 600 mile trip	Open-ended	13.0	2.9	7
#2 2400 gallon tank	Goal specification	13.2	2.2	8
#3 720 pages	Equation setup	13.3	2.8	7
#4 DOT road crew	Faulty solution	12.4	4.4	10

Within problem class, items were generally of similar difficulty. The single notable exception was the three-goal "active ingredient" problem, which was substantially harder than its counterparts. This difference, reflected in the variance components analysis as a large format effect, is possibly due to the need to manipulate the decimal percentage, .25%. The range of scores within problem classes varied widely, with problems like the two-goal "small business" item and the three-goal "graphics designer" question displaying high restriction and others showing a more acceptable dispersion. These dispersion differences might be expected to produce some degree of artifactual variation in the product-moment correlations between GIDE and the content experts.

Also complicating the assessment of agreement is the high frequency of perfect responses for some problems. Perfect responses should be somewhat easier for both humans and GIDE to evaluate because the potential set of answers is more constrained. As a result, agreement for these responses should be higher than for imperfect ones.[6]

Table 4 presents the Pearson product-moment correlations between the mean score for an item (taken across raters for each examinee) and GIDE's rating for that item. These values are based on the complete sample, as GIDE was able to produce a score for each of the 360 cases (12 items x 30 examinees). The median of the 12 correlations is .88, with most values ranging from the middle eighties to middle nineties, suggesting that GIDE's rank ordering of examinees is highly similar to that of human judges'. The only value falling well outside this range, .74, resulted almost entirely from two large score discrepancies. The within-class medians--.87 for the two-goal problems, .85 for three-goal, and .91 for five goal--are reasonably similar to one another, as are the medians for item format (.89 for open-ended, .85 for goal specification, .92 for equation setup, and .86 for faulty solution).

The mean scores awarded by the raters and by GIDE are shown in Table 5. These data give an indication of the extent to which GIDE is generally too easy or hard in its grading. As the table suggests, GIDE's scores are in most cases lower and more dispersed than the content experts' scores. This tendency occurs across problem class and format.

Table 6 presents the distribution of discrepancies between GIDE's scores and the mean of the raters' scores for each problem within a set, where a discrepancy is calculated by subtracting GIDE's score from the rater mean. To summarize each distribution, the mean absolute and mean signed differences are given, the former indicating how far off GIDE's scores are on average and the latter giving the direction and magnitude of any systematic bias. This latter indicator is also shown scaled in standard deviation units as the signed/SD ratio, the mean signed

Table 4

Product-Moment Correlations Between Raters' Mean Scores and GIDE's Scores for Items within Problem Sets (N=30)

Problem Set & Item	Item Format	Product-Moment Correlation
Two-goal		
#1 2000 cc tank	Open-ended	.89
#2 Small business	Goal specification	.85
#3 Chemical company	Equation setup	.95
#4 $3.50 in tolls	Faulty-solution	.74
Median		.87
Three-goal		
#1 Investment fund	Open-ended	.83
#2 Load of cement	Goal specification	.93
#3 Graphics designer	Equation setup	.83
#4 Active ingredient	Faulty-solution	.86
Median		.85
Five-goal		
#1 600 mile trip	Open-ended	.90
#2 2400 gallon tank	Goal specification	.84
#3 720 pages	Equation setup	.92
#4 DOT road crew	Faulty-solution	.97
Median		.91

Table 5

Means and Standard Deviations of Scores Given by GIDE and Raters

Two-Goal Problems (scale = 0-6)						
		Rater				
Problem	GIDE	#1	#2	#3	#4	#5
#1 2000 cc tank						
Mean	4.8	5.1	4.9	5.2	5.1	5.0
SD	2.3	1.8	2.1	1.9	1.8	2.0
#2 Small business						
Mean	5.4	5.7	5.6	5.6	5.7	5.7
SD	1.3	.5	.6	.8	.7	.5
#3 Chemical company						
Mean	5.3	5.4	5.4	5.4	5.6	5.6
SD	1.6	1.5	1.5	1.5	1.0	1.2
#4 $3.50 in tolls						
Mean	5.6	5.3	5.6	5.5	5.6	5.6
SD	1.0	1.5	1.0	1.2	1.0	.9

Three-Goal Problems (scale = 0-9)						
		Rater				
Problem	GIDE	#1	#2	#3	#4	#5
#1 Investment fund						
Mean	7.7	8.3	8.2	8.3	8.2	8.3
SD	2.5	2.0	2.1	2.0	1.9	2.0
#2 Load of cement						
Mean	8.3	8.6	8.6	8.6	8.5	8.5
SD	1.6	1.2	1.1	1.1	1.3	1.5
#3 Graphics designer						
Mean	8.5	8.6	8.8	8.9	8.9	8.9
SD	1.6	1.1	.6	.3	.4	.3
#4 Active ingredient						
Mean	5.8	4.9	6.0	6.0	5.8	5.7
SD	2.8	3.5	2.5	2.8	2.9	2.9

Five-Goal Problems (scale = 0-15)						
		Rater				
Problem	GIDE	#1	#2	#3	#4	#5
#1 600 mile trip						
Mean	12.8	13.1	12.8	13.1	12.8	13.2
SD	2.9	2.9	2.9	2.9	3.1	2.7
#2 2400 gallon tank						
Mean	12.1	13.3	13.2	13.0	13.3	13.1
SD	3.3	2.1	2.2	2.4	2.1	2.2
#3 720 pages						
Mean	12.9	13.4	13.2	13.4	13.2	13.1
SD	2.9	2.8	2.8	2.8	2.8	2.8
#4 DOT road crew						
Mean	11.8	12.5	12.2	12.2	12.7	12.5
SD	4.7	4.7	4.7	4.8	4.0	4.0

Table 6

Frequency Distribution of Differences Between GIDE's Scores and Judges' Mean Ratings by Problem within Set (N = 30)

Two-Goal Problems (scale = 0-6)				
	Problem			
Difference	#1	#2	#3	#4
> 6.0				
5.0 to 5.9	1			
4.0 to 4.9		1		
3.0 to 3.9	1			
2.0 to 2.9			1	1
1.0 to 1.9		1	2	
.1 to .9	2	5	2	
0.0	24	22	22	23
- .1 to - .9	2	1	3	5
-1.0 to -1.9				
-2.0 to -2.9				
-3.0 to -3.9				1
-4.0 to -4.9				
-5.0 to -5.9				
<-6.0				
Mean absolute	.3	.3	.2	.2
Mean signed	.3	.2	.2	-.1
Signed/SD ratio	.1	.3	.1	-.1

Three-Goal Problems (scale = 0-9)				
	Problem			
Difference	#1	#2	#3	#4
> 5.0				
4.0 to 4.9	2		2	1
3.0 to 3.9	2	1		
2.0 to 2.9	2			1
1.0 to 1.9		3		1
.1 to .9	1			3
0.0	21	26	27	16
- .1 to - .9	1		1	3
-1.0 to -1.9	1			1
-2.0 to -2.9				2
-3.0 to -3.9				2
-4.0 to -4.9				
<-5.0				
Mean absolute	.7	.2	.3	.8
Mean signed	.5	.2	.3	-.1
Signed/SD ratio	.3	.2	.4	.0

Note. Positive differences indicate that the judges' mean score was higher than GIDE's score. The signed/SD ratio is the ratio of the mean signed difference t the mean standard deviation taken across all raters and GIDE.

Table 6 (con't)

Frequency Distribution of Differences Between GIDE's Scores and Judges' Mean Ratings by Problem within Set (N = 30)

Five-Goal Problems (scale = 0-15)				
	Problem			
Difference	#1	#2	#3	#4
> 7.0				
6.0 to 6.9		1		
5.0 to 5.9		1	1	1
4.0 to 4.9	1	2		
3.0 to 3.9		2		
2.0 to 2.9	4	1	2	2
1.0 to 1.9	1	2	2	6
.1 to .9	2	4	4	5
0.0	13	11	12	14
- .1 to - .9	6	5	8	1
-1.0 to -1.9	2	1	1	1
-2.0 to -2.9	1			
-3.0 to -3.9				
-4.0 to -4.9				
-5.0 to -5.9				
-6.0 to -6.9				
<-7.0				
Mean absolute	.7	1.2	.6	.8
Mean signed	.3	1.0	.3	.6
Signed/SD ratio	.1	.4	.1	.1

Note. Positive differences indicate that the judges' mean score was higher than GIDE's score. The signed/SD ratio is the ratio of the mean signed difference to the mean standard deviation taken across all raters and GIDE.

difference divided by the mean of the six score standard deviations for an item (five raters plus GIDE). Though mostly positive, the mean signed differences are generally small, indicating a trivial bias (a tenth of a standard deviation unit or less) in 7 of the 12 cases. Mean signed differences of .3 to .4 standard deviations appear in 4 cases and show no particular association with problem class or format. In absolute terms, even these discrepancies are quite small, appearing consequential in standard deviation units largely because of the restricted range of examinee scores. The largest mean absolute differences in each problem class were .3, .8, and 1.2 points for the two-, three-, and five-goal problems, respectively. These discrepancies translate to 4%, 8%, and 8% of the range of the respective 7-, 10-, and 16-point score scales.

As the distributions themselves confirm, the overwhelming majority of discrepancies were relatively small. Still, there were several quite substantial deviations. For the two-goal problems, 6 (of 120) responses had discrepancies of two or more points on the seven-point score scale. Ten had differences of three points or larger for the three-goal problems (graded on a ten-point scale). For the five-goal problems, scored on a 16-point scale, GIDE's scores for seven responses were discrepant by four or more points.

Presented in Table 7 are the mean discrepancies for perfect versus imperfect responses, where the distinction is based on the mean of the raters' scores. Because perfect responses might be easier for the machine and raters to agree upon, the discrepancies for the imperfect papers might provide a more conservative measure of GIDE's accuracy. As the table shows, for 9 of the 12 items GIDE and the raters agree completely in scoring perfect responses. Where there is less than perfect agreement, it is trivial; the largest mean absolute discrepancy is .6 on a 16-point scale. For imperfect responses, the values are considerably higher--and surprisingly uniform--with medians of 1.0, 1.2, and 1.0 for the three problem sets, respectively. The single noticeably discrepant value occurs for the three-goal equation setup problem, for which the mean absolute discrepancy is 2.4 (on a ten-point scale). This mean is, however, based on only four imperfect responses and may not be dependable.

What causes might underlie the few large discrepancies detected? To address this question, GIDE's processing of the 23 large discrepancies cited above was analyzed (6 responses to the two-goal problems, 10 three-goal responses, and 7 five-goal ones). Four general sources were discovered.

The single most frequent cause was transcription error. This cause accounted for 9 of the 23 large discrepancies and affected 6 of the 12 problems (especially the five-goal ones, which required the longest transcriptions.) These human errors most often involved changing a value from the original to the copy or leaving out essential steps. When the 9 errors were

Table 7

Mean Absolute Discrepancies Between Rater Mean Scores and GIDE Scores for Perfect and Imperfect Responses (N = 30)

Problem Set & Item	Item Format	Perfect Response	Imperfect Response
Two-goal (scale = 0-6)			
#1 2000 cc tank	Open-ended	.0 (22)[a]	1.2 (8)
#2 Small business	Goal specification	.0 (21)	.8 (9)
#3 Chemical company	Equation setup	.0 (21)	.8 (9)
#4 $3.50 in tolls	Faulty-solution	.0 (23)	1.1 (7)
Median		.0	1.0
Three-goal (scale = 0-9)			
#1 Investment fund	Open-ended	.0 (24)	1.2 (6)
#2 Load of cement	Goal specification	.0 (23)	1.0 (7)
#3 Graphics designer	Equation setup	.0 (26)	2.4 (4)
#4 Active ingredient	Faulty-solution	.0 (9)	1.2 (21)
Median		.0	1.2
Five-goal (scale = 0-15)			
#1 600 mile trip	Open-ended	.0 (7)	1.0 (23)
#2 2400 gallon tank	Goal specification	.6 (8)	1.4 (22)
#3 720 pages	Equation setup	.3 (7)	.7 (23)
#4 DOT road crew	Faulty-solution	.2 (10)	1.0 (20)
Median		.3	1.0

Note. Perfect papers were defined as those for which the mean score taken across human raters was the maximum score that could be received.

[a]The number of responses appears in parentheses.

corrected, the two-goal problem median changed from .87 to .91, the three-goal median from .85 to .86, and the five-goal value from .91 to .94. The median across all 12 problems rose from .88 to .89.

A second cause of discrepancy involved several correctable difficulties with GIDE's inference mechanism. The most significant difficulty appeared to account for GIDE's general tendency to award slightly lower scores than the raters. The appropriate solutions to several problems include values that are close to one another. For example, the "600 mile" problem has one distance segment traveled in 6.3 hours and another in 6.33 hours. For scoring and error diagnosis to be implemented properly, GIDE needs to associate each of these values with the appropriate goal. One way it does this association is by imposing value constraints, that is, comparing incorrect values to a range of "reasonableness" built around the expected result. In general, these ranges tended to be more restrictive than those applied by the raters, sometimes causing GIDE to infer that a goal was missing or structurally incorrect when it was more properly considered a computational error. These overly strong value constraints accounted for five large discrepancies among the interest problems.

Another inferencing difficulty pertained to unit labels. Some students always use labels, some leave them off entirely, and some use them inconsistently within a given problem. When units are present in the solution, GIDE uses them in its evaluation. When not explicitly stated, GIDE assumes the student is using the units specified by the problem stem (unless the solution presents evidence to the contrary). This strategy works well except when students work the problem correctly in different units (e.g., minutes instead of hours) and fail to use labels consistently. In these instances, GIDE correctly processes the stated labels when attached to specific values, but switches to the problem-specified units when labels are absent, causing some pieces of the solution to be erroneously interpreted.

The third major cause of discrepancy concerned GIDE's knowledge bases. In two cases, student errors fell outside the faulty plans to which GIDE had access. (Because the errors were consistent with GIDE's knowledge of the solution structure, the plans can be easily added.) Two other cases represented partially correct solutions taking novel problem-solving approaches, which the raters were able to decipher. GIDE can appropriately score partially correct responses that follow solution paths similar to ones in its knowledge base. It can also handle novel correct responses by checking the accuracy of the end result. It cannot, however, easily handle responses that are both novel and partially correct, because it has no comparable solution structure to use as a reference.

The fourth major cause of discrepancy was differences between GIDE and the raters in applying the scoring rubric.

Generally, GIDE's application appeared defensible, if not completely correct. On one problem, GIDE and the raters disagreed legitimately on the classification of a particular error. For another, the raters split into two groups with each group assigning dissimilar scores. GIDE closely agreed with one group but, because of the difference between the groups, diverged from the rater mean. In a third instance, GIDE's dissonance with the judge mean came about solely because of a single judge's recording error (awarding a perfect score to a meaningless response).

Discussion

This study assessed agreement between expert system and human judges' partial-credit scores on complex constructed-response algebra word problems adapted from the GRE General Test. Problems were drawn from three content classes and four constructed-response formats. Results showed the expert system to produce scores for all responses and to duplicate the raters' judgments with reasonable accuracy. The program achieved a median correlation across 12 problems of .88 and widely diverged from human judgments in only a small number of cases. Moreover, the root causes of these few divergences were either irrelevant to evaluating the system's scoring accuracy (i.e., transcription errors), or largely correctable, involving such things as improvements to the inference mechanism, additions of faulty plans to the knowledge base, or clarifications of the rubric.

In considering these results, it is instructive to compare GIDE's performance with MicroPROUST (Johnson & Soloway, 1985; Johnson, 1986), a related expert system that scores responses to computer programming problems. For two versions of a faulty solution item, MicroPROUST was able to score 82% and 85% of the responses presented it. For these responses, its scores correlated .82 and .88 with a single human rater (Braun et al., 1990). For two open-ended problems, only 72% of the responses used in developing the knowledge base could be evaluated. This figure dropped to 42% in an independent sample. Correlations between the program and the mean of four raters' scores for the initial response set were .75 and .96 (Bennett, Gong, Kershaw, Rock, Soloway, & Macalalad, 1990).

The disparity between the two systems in generating scores is a reflection both of differences in the content domains and of how each program was adapted to those differences. In analyzing algebra solutions, GIDE is able to use numerical values as well as symbolic structures. Numerical values provide a means of evaluation not available in the programming domain. In addition, unlike programs, algebra solutions are often structured such that the realization of later goals presumes that earlier ones have been achieved. This hierarchical ordering permits GIDE to infer that some steps have been completed even if they have not been explicitly stated or are stated in a way that GIDE cannot

understand. This structure might help explain GIDE's successful performance with both open-ended and constrained formats.

MicroPROUST has no such interpretive advantage. To keep its scoring reasonably accurate, it was programmed to act conservatively, withholding judgment on solutions containing segments it could not interpret. This strategy limits the incidence of large scoring errors, but at the cost of reducing the proportion of solutions the program will grade. Its performance is more sensitive to item format, degrading considerably when encountering open-ended questions.

Although GIDE's scores would appear to closely approximate those of human graders, they are not interchangeable. Experts possess a depth of domain understanding, a network of related knowledge, and flexibility in applying both that, in concert, help guide scoring decisions. As this study implies, these characteristics might be partially duplicated in a computer program such that most scores would be comparable to an expert's. However, at least for the foreseeable future, it is unlikely that these characteristics could be fully represented. Consequently, some responses--particularly those containing creative but slightly incorrect approaches--will not be scored equivalently.

As noted, GIDE is currently a batch processing laboratory tool. Of the potential directions its development could take, two merit particular discussion. The use that fits best with GIDE's current state is a grader's assistant. In this application, GIDE would centrally batch-process answers to constructed-response problems administered in remote computer-based environments. Human judges would play a critical role in this scenario, as they do in such programs as Advanced Placement (College Entrance Examination Board, 1988). But instead of taking primary responsibility for scoring, the content expert's job would be to verify the accuracy of GIDE's analyses. To do this, the expert would rapidly review on a computer monitor a version of the examinee's solution on which were graphically indicated the location, type, and scoring of each error GIDE detected. If the expert disagreed with GIDE's analysis, he or she would electronically note the disagreement directly on the solution, which would then be routed to a second judge for arbitration. If experience showed disagreement to be associated with particular response characteristics, selected responses might be automatically identified and human verification restricted to this subset.

Perhaps a more interesting direction is toward producing an interactive, constructed-response, diagnostic assessment system. The research and development needed to build a strong foundation would seem to require psychometric, test development, and technological components. From a psychometric perspective, there are several important questions to answer about the meaning of GIDE's numeric scores. One central issue is construct validity. In particular, evidence needs to be gathered on whether the

constructed-response item types measure a dimension different from GRE-quantitative items, and, if so, how that ability is different--in short, what formats should compose a diagnostic system and how that system should be positioned relative to the GRE-quantitative section. High construct overlap might suggest a diagnostic assessment taken in preparation for the General Test. Low overlap, accompanied by higher relations with important cognitive criteria, might signal a potential problem-solving complement to the quantitative section, perhaps the precursor to a new General Test scale or other new program offering.

A second psychometric issue concerns refining GIDE's scoring rubric. As indicated by the analysis of large score discrepancies, the definitions of the error classes need sharpening. An additional objective might be a modification that places all items on the same partial-credit scale but retains the strong domain links characteristic of the current scheme. A single score scale obviously makes it easier to compare the functioning of items, perform item-level analyses of group performance, and uniformly apply certain statistical methods. Grounding the rubric in the domain analysis (i.e., framing it around the way students and experts actually solve the problems) is similarly valuable because it gives scores deeper meaning and permits them to be closely tied to error diagnoses.

Diagnostic analysis represents a third component of the psychometric foundation of any functional system. The impressive scoring agreement reported here provides support for the general soundness of GIDE's item-level error diagnoses because its scores are generated directly from these judgments: to produce accurate scores, GIDE must have found numbers and types of errors similar to what the human judges found. Whether its judgments are correct about the <u>specific</u> nature of these errors is another matter. It is possible for GIDE and the raters to agree, for example, that a structural error is present and--by reference to the rubric--deduct the same number of points, yet still describe the error differently. Such fine distinctions, while unimportant for generating numeric scores, are surely critical in providing accurate feedback about why an individual received a particular score and what confusions that person's solution evidenced.

Though item-level diagnoses can provide useful information about exactly what errors were made in a response, more dependable, general, and arguably more powerful feedback might be generated at the test level. The foundational psychometric element needed regards how best to aggregate information across items so as to detect and characterize instructionally meaningful patterns in the numeric score or error diagnosis data. Probabilistic models for cognitive diagnosis have only recently become available (Mislevy, in press; Masters & Mislevy, in press), and it is not yet clear what models or diagnostic characterizations might best suit given assessment purposes. As such, it is likely that additional models and characterizations will need to be studied.

From a test development perspective, the main thrust should be increasing domain coverage. One way is to deepen coverage by adding isomorphic items to the existing problem classes. Such additions are very efficient because the costly knowledge bases needed to analyze responses already exist. Variation in difficulty might be achieved through items that are essentially parallel in solution process but that contain values more difficult to manipulate (e.g., as apparently occurred in the "active ingredient" problem). A second approach is to gradually increase breadth through new item classes and, eventually, new General Test mathematical domains (e.g., geometry and arithmetic). A first step might be to build out from the existing core of rate, work, and interest problems to such closely related content classes as graduated rate and percent change.

Several technological components would be needed for an operational system. Perhaps the most costly additions would be the knowledge bases to support content expansion. How much effort might be required? As suggested, adding isomorphs is relatively simple and should involve only a day of knowledge base development per item. Inserting problems from related content classes (e.g., graduated rate, percent change) is considerably more involved, probably requiring two person-months per class. The most labor-intensive task would be developing the capability to analyze problems from several content classes belonging to another mathematics domain. This task is comparable to the one undertaken in the current project. For this project, 70 person-days were devoted to developing knowledge bases for the three problem classes; a comparable investment was associated with the related task of improving GIDE's inference mechanism and of adapting it from analyzing statistics items to algebra word problems.

These tasks are unarguably labor intensive: it would take an enormous effort to cover a domain the size of, for example, high school mathematics. These costs become more reasonable, however, when viewed from the perspective of infrastructure development (e.g., consider the 50 or so years devoted to this task for multiple-choice tests). This perspective suggests that tools might be built to make development more efficient and that economies of scale might be realized. For example, once knowledge bases are built, they can be used to analyze responses to any item written to a given specification. Additionally, existing knowledge bases should be useable as components of knowledge bases for items that test related skills. Knowledge bases can also serve as the basis of systems for teaching students to solve the same problems used in assessment--expert approaches to problem solution compose the knowledge base, as do common errors. Finally, the knowledge bases can be employed to help test developers write multiple-choice questions whose distractors better capture common conceptually salient errors, something that current multiple-choice tests might not

effectively do (Bridgeman, personal communication, July 12, 1990).

A second high technological priority is an interactive interface. This development is necessary not only to avoid the trouble and expense associated with data transcription but, more importantly, to bring data collection conditions closer to the interactive testing environment to which results must be generalized.

In designing an interface, significant consideration needs to be given to response constraint. Enough constraint should be imposed to permit responses to be analyzed without human editing, but not so much as to distort the response process. One example might be the ability to process the natural language responses examinees occasionally pose. These responses are typically associated with a simulation approach to problem solving (e.g., "If the net water level in the tank increases by 16 cc each minute, then after 100 minutes it will contain 1600 cc, so that to fill the 2,000 cc tank would take another 25% or 125 minutes."). Where feasible, automatically translating such simulations to equation form seems preferable to forcing changes in examinee solution strategy.

A longer term technological requirement is redesigning GIDE's inference mechanism for an operational setting. This redesign should correct those failings discovered in the present study (i.e., overly strict value constraints, handling of unit labels), as well as ones discovered in any more systematic investigation of GIDE's diagnostic accuracy. In addition, the task would involve optimizing GIDE's matching algorithm and coding it in a production language (e.g., "C") for increased efficiency. In concert with this redesign, modifications to the knowledge representation might be considered to capture novel problem-solving approaches better and to make the knowledge bases more compact. Last, tools might be built to make knowledge base development less labor intensive. One possibility here is to automate the transfer of knowledge from the mathematical representations produced by a content expert to machine-readable form.

Several limitations of the current study should be noted. The first regards the assignment of transcribed solutions to GIDE versus original productions to the raters. This differential assignment was done to tie estimates of GIDE's scoring accuracy as closely as possible to the solutions produced by students--which, for obvious reasons, GIDE could not directly rate. Having raters grade only the transcripts would have produced agreement estimates with limited generalizability to original productions. A design in which overlapping groups of raters graded original and transcribed solutions might have provided an efficacious solution, but resource constraints prevented this approach. The effect of differential assignment on estimating agreement is not immediately evident. The original solutions might have given the

experts more detailed information than could be represented in the copies (e.g., diagrams, verbal notes). On the other hand, the originals might have been more ambiguous than the linearized transcriptions. GIDE might have benefited from some degree of clarification unintentionally imposed by the transcribers, or it might have been thrown off (as the evidence suggests it sometimes was) by transcription errors. In any event, an interactive prototype would remove these extraneous influences by capturing a single production that both GIDE and the raters can assess.

A second limitation concerns the number of algebra problems and content classes employed. Combining instrument development and empirical research in a single investigation prevented a more comprehensive item production effort. New isomorphs and problems from different content classes will eventually broaden the universe over which GIDE's performance can be generalized, as will problems covering a wider range of difficulty.

Finally, the nonrepresentative nature and small size of the study sample should be noted. The sample was somewhat more mathematically adept than the test-taking population, a factor that probably contributed to the restricted range of scores on some problems. These restricted ranges might also have been caused by more liberal time limits than those offered on the General Test. In future work, greater efforts might be made to attract less skilled segments of the General Test population and to impose stricter timing constraints.

Together with the generally positive findings from the computer science domain (Bennett, Gong, Kershaw, Rock, Soloway, & Macalalad, 1990; Bennett, Rock, Braun, Frye, Spohrer, & Soloway, 1990; Braun et al., 1990), these results suggest considerable promise for new approaches to standardized assessment. In particular, these findings move us closer to systems that present problems similar to those encountered in academic and work settings, that capture varying degrees of solution correctness, and that recognize, describe, and perhaps help remediate the errors examinees make.

References

Bennett, R. E. (In press). Toward intelligent assessment: An integration of constructed response testing, artificial intelligence, and model-based measurement. In N. Frederiksen, R. J. Mislevy, and I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E., Gong, B., Kershaw, R. C., Rock, D. A., Soloway, E., & Macalalad, A. (1990). Assessment of an expert system's ability to grade and diagnose automatically student's constructed responses to computer science problems. In R. O. Freedle (Ed.), Artificial intelligence and the future of testing (pp. 293-320). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bennett, R. E., Rock, D. A., Braun, H. I., Frye, D., Spohrer, J. C. & Soloway, E. (1990). The relationship of constrained free-response to multiple-choice and open-ended items. Applied Psychological Measurement, 14, 151-162.

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). Toward a framework for constructed-response items (ETS RR-90-7). Princeton, NJ: Educational Testing Service.

Braun, H. I., Bennett, R. E., Frye, D, & Soloway, E. (1990). Scoring constructed responses using expert systems. Journal of Educational Measurement, 27, 93-108.

College Entrance Examination Board. (1988). Technical manual for the Advanced Placement Program. New York: Author.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons.

Fiske, E. B. (1990, January 31). But is the child learning? Schools trying new tests. The New York Times, pp. 1, B6.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27-32.

Guthrie, J. T. (1984). Testing higher level skills. Journal of Reading, 28, 188-190.

Johnson, W. L. (1986). Intention-based diagnosis of novice programming errors. Los Altos, CA: Morgan Kaufmann Publishers.

Johnson, W. L., & Soloway, E. (1985). PROUST: An automatic debugger for Pascal programs. Byte, 10(4), 179-190.

Masters, G. N., & Mislevy, R. J. (In press). New views of student learning: Implications for educational measurement. In N. Frederiksen, R. J. Mislevy, and I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Matz, M. (1982). Towards a process model for high school algebra. In D. H. Sleeman and J. S. Brown (Eds.), Intelligent tutoring systems. London: Academic Press, Inc.

Mislevy, R. J. (In press). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, and I. Bejar (Eds.), Test theory for a new generation of tests. Hillsdale, NJ: Lawrence Erlbaum Associates.

Nickerson, R. S. (1989). New directions in educational assessment. Educational Researcher, 18(9), 3-7.

Reed, S. K. (1987). A structure-mapping model for word problems. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 124-139.

Reed, S. K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving algebra word problems. Journal of Experimental Psychology: Learning, Memory, and Cognition, 11, 106-125.

Sebrechts, M. M., LaClaire, L., Schooler, L. J., & Soloway, E. (1986). Toward generalized intention-based diagnosis: GIDE. In R. C. Ryan (Ed.), Proceedings of the 7th National Educational Computing Conference (pp. 237-242). Eugene, OR: International Council on Computers in Education.

Sebrechts, M. M., & Schooler, L. J. (1987). Diagnosing errors in statistical problem solving: Associative problem recognition and plan-based error detection. Proceedings of the Ninth Annual Cognitive Science Meeting (pp. 691-703). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sebrechts, M. M., Schooler, L. J., LaClaire, L., & Soloway, E. (1987). Computer-based interpretation of students' statistical errors: A preliminary empirical analysis of GIDE. Proceedings of the 8th National Educational Computing Conference (pp. 143-148). Eugene, OR: International Council on Computers in Education.

Thorndike, R. L. (1982). Applied psychometrics. Boston: Houghton Mifflin.

Footnotes

1. Field office locations were Atlanta, GA; Austin, TX; Brookline, MA; Emeryville, CA; Evanston, IL; Pasadena, CA; Princeton, NJ; and Washington, D.C.

2. The goal decomposition for a problem is derived from the cognitive analysis of expert and novice solutions described above.

3. Any eventual operational application would obviously need to be computer based if performance information were to be provided at the time of testing.

4. Four of these individuals also contributed problem solutions for the knowledge base and advised on the scoring rubric. Because the study addressed whether, given a common rubric and overlapping knowledge base, a machine could duplicate the judgments of human content experts, this dual use was not viewed as problematic.

5. Though the size of this component might have been due to administration order effects, post-hoc examination of the item means for each order failed to substantiate this hypothesis.

6. This is not to suggest that scoring a perfect response is necessarily straightforward. GIDE treats all solutions equivalently, attempting to construct a step-by-step understanding of the response process, as opposed to basing its scores only on the correctness of the final result. (The raters should act similarly, given the directions of the scoring rubric.) Constructing a step-by-step understanding is nontrivial, as there are multiple ways to arrive at the same result. For all solutions, then, scoring is an analytical process.

Appendix A

Constructed-Response Item Stems

Work Prototype (Two-goal problems)

How many minutes will it take to fill a 2,000-cubic-centimeter tank if water flows in at the rate of 20 cubic centimeters per minute and is pumped out at the rate of 4 cubic centimeters per minute? (OE)

Isomorphs

One of two outlets of a small business is losing $500 per month while the other is making a profit of $1750 per month. In how many months will the net profit of the small business be $35,000? (GS)

A specialty chemical company has patented a chemical process that involves 2 reactions. Reaction 1 generates 24 grams of molecule B per minute and reaction 2 consumes 5 grams of molecule B per minute. If 4,560 grams of molecule B are desired as a product of this process, how many minutes must it continue? (ES)

$3.50 in tolls is received each minute at an automated toll booth while the rate at a booth with an operator is $2.80 each minute. How many minutes elapse before the automated booth receives $14.00 more in tolls than does the person-operated booth? (FS)

Interest Prototype (Three-goal problems)

Money in a certain investment fund earns an annual dividend of 5 percent of the original investment. In how many years will an initial investment of $750 earn total dividends equal to the original investment? (OE)

Isomorphs

On every $150 load of cement it delivers to a construction site, Acme Cement Company earns a 4 percent profit. How many loads must it deliver to the site to earn $150 in profit? (GS)

A graphics designer earns 2% of a $1500 yearly bonus for each shift of overtime she works. How many shifts of overtime must she work to earn the equivalent of the entire yearly bonus? (ES)

The active ingredient is 0.25 percent of a 3-ounce dose of a certain cold remedy. What is the number of doses a patient must take before receiving the full 3 ounces of the active ingredient? (FS)

Rate x Time Prototype (Five-goal problems)

On a 600-hundred mile motor trip, Bill averaged 45 miles per hour for the first 285 miles and 50 miles per hour for the remainder of the trip. If he started at 7:00 a.m., at what time did he finish the trip (to the nearest minute)? (OE)

Isomorphs

800 gallons of a 2,400 gallon tank flow in at the rate of 75 gallons per hour through a clogged hose. After the hose is unclogged, the rest of the tank is filled at the rate of 250 gallons per hour. At what time to the nearest minute will the filling of the tank be finished if it starts at 5:30 a.m.? (GS)

Of the 720 pages of printed output of a certain program, 305 pages are printed on a printer that prints 15 pages per minute and the rest are printed on a printer that prints at 50 pages per minute. If the printers run one after the other and printing starts at 10 minutes and 15 seconds after the hour, at what time to the nearest second after the hour will the printing be finished? (ES)

A Department of Transportation road crew paves the 15 mile city portion of a 37.4 mile route at the rate of 1.8 miles per day and paves the rest of the route, which is outside the city, at a rate of 2.1 miles per day. If the Department of Transportation starts the project on day 11 of its work calendar, on what day of its work calendar will the project be completed? (FS)

Note. The format in which the item was presented is indicated in parentheses after each item. OE = open-ended, GS = goal specification, ES = equation setup, FS = faulty solution.

Appendix B

Transcription Rules and Examples of Original and Transcribed Responses

RULES FOR TRANSCRIBING WRITTEN STUDENT SOLUTIONS TO MACHINE SCORABLE FORMAT
V 1.3 3/11/90

The general goal in transcription is to write out the students solution in a linear fashion that closely approximates the student's solution. As much of the solution is included as possible, with an attempt to replicate the order used by the student. The main difference in the machine and student solutions will be:
(1) Lines are written in sequence instead of being laid out spatially on a page.
(2) Solutions are cleaned up syntactically so that they can be parsed.
(3) Illegible or irrelevant portions of solutions (e.g. intermediate products) are ignored.

GENERAL FORMAT
1. Follow the sequence in the students' solutions insofar as possible.
2. For items without specified sequence group related equations.
3. All solutions include a common structure consisting of the following (Both upper and lowercase are acceptable, but we will adopt the convention of using uppercase to improve readability.):

GIVENS:
UNKNOWNS:
ANSWER =

4. Labels for GIVENS and UNKNOWNS that are presented as part of the problem statement and have associated lines for students' solution steps are included in the solution if the student has written some part of the solution on the associated lines. If there is no value associated with the answer, write ANSWER = 99999. For all other labels which have no associated terms, drop the label.

Eg: GIVENS:
Percent_Profit = 4%
Cost_per_Load = $ 150
Target_Profit = $150
UNKNOWNS:
Profit_per_Load = 4 % * $ 150
4 % * $ 150 = $ 6
Loads_Needed_for_Target_Profit = Target_Profit - Profit_per_Load
Target_Profit - Profit_per_Load = 150 / 6
150 / 6 = 25 loads
ANSWER = 25 loads

5. Each line should be written as a SYNTACTICALLY correct equation.
6. Only ONE equation is allowed per line; if multiple equalities occur on a single line, put each equality as a separate equation on its own line.

E.g. Profit_per_Load = 4 % * $ 150 = $ 6 =-> Profit_per_load = 4 % * $ 150;
4 % * $ 150 = $ 6

7. Disregard incomplete or illegible statements or computations
8. If expressions under the Calculations header are redundant with those on the line, only include the expression on the line.

NUMBERS
1. Numbers should be integer or real.
2. No numbers shall begin or end in a decimal point (e.g. "0.6" NOT ".6"; "6" or "6.0" NOT "6.")
3. Repeating decimals should be represented to the nearest hundredth.
4. Positive numbers should be written without a sign (+10 =-> 10); negative numbers should include the sign (-10).

UNITS

1. Units should be written in a standard form, either with a name (5 hrs; 10 dollars) or with a specification of rate (5 miles per hour; 4 dollars per month).
2. If a unit appears without a quantity assume 1 unit. E.g. hr = 1 hr
3. Expressions of complex ratio units should be written out. (Operators should NOT be included in units.)
   E.g. 5 miles / hr =-> 5 mph or 5 miles per hour
4. Incomplete units should be completed when interpretable or dropped if uninterpretable.
   E.g. r1 = 45 per hr =-> r1 = 45 mi per hr
5. Include any units that are NECESSARY for problem interpretation and are UNAMBIGUOUS.
   E.g. On problems requiring time determination: 7 - 7 = 12 =-> 7 p.m. - 7 a.m. = 12 hrs

TIME

1. Elapsed time is given as hrs and minutes in the form "M hrs B mins"; no "and" should be used in presenting clock time.
   E.g. 10 hrs and 36 minutes =-> 10 hrs 36 minutes
2. Clock time may be represented in its standard civilian form (e.g. 10:36). Conventions for before and after noon should be one of the following a.m.; am; p.m.; pm. In the absence of a.m. or p.m., a.m. is assumed. Military time (e.g. 15:42) is also fine.

SPACING

1. Units, values, and operators are separated from each other by spaces.
   (e.g. 5 % NOT 5%; 285 miles / 45 miles per hour = 6 1/3 hours NOT 285miles/45milesperhour=61/3hours; $ 150 NOT $150)
2. Eliminate unnecessary punctuation, including commas.
   ($ 35,000 =-> $ 35000; but 10:38 remains 10:38).
3. In the special case of NEGATIVE $, the $ should come first (-$300 =-> $ -300).

OPERATORS

1. If a word is used in place of the operator, replace the word with the associated operator to form a syntactically correct equation.
   (E.g. Dividend is 5 % of $750 =-> Dividend = 5 % * $ 750)
2. Operators CANNOT be used in LABELS. E.g. t1_+_t2 =-> t1 + t2 or t1_and_t2

SPECIAL CHARACTERS

1. No special characters are allowed; they should be converted to a word equivalent.
   (e.g. #_of_doses =-> no_of_doses)
2. Isolated question marks should be replaced by variable names, X1, X2, etc.

LABELS

1. If the spacing in a label is ambiguous, use a consistent spacing for multiple ambiguities.
   (e.g. Filling_Amount_1 vs. Filling_Amount1)
2. Eliminate surface structure differences for common labels. Make subsequent labels conform to the structure of the first instance of its use.
   E.g. Fill_Time_1 = 800 / 75
   Filling_Tim_1 = 10 2/3 =-> Fill_Time_1 = 10 2/3
2. For unlabeled values, assign that value to the most proximate label.
   E.g. Filling_Amount_2 = Tank_Cap - Filling_Amount_1
   1600 gals =-> Filling_Amount_2 = 1600 gals
3. Unlabeled numbers for which no proximate label is available should be given a label X#, where the number begins with 1 and increments to the next integer for the next unlabeled value (e.g. 754 loads =-> X1 = 7 loads).
4. Labels may not begin with special characters such as "$" or "?" or with numbers.

ANSWER
1. The answer should be written in a standard form as ANSWER = # UNIT. Other extraneous words should be eliminated. (E.g. ANSWER = completed 29 =-> ANSWER = 29)

261

PLEASE PRINT YOUR SOLUTION CLEARLY. Name: [redacted]

Problem 12. Of the 720 pages of printed output of a certain program, 305 pages are printed on a printer that prints 15 pages per minute and the rest are printed on a printer that prints at 50 pages per minute. If the printers run one after the other and printing starts at 10 minutes and 15 seconds after the hour, at what time to the nearest second after the hour will the printing be finished?

Equations that Will Provide a Solution:

Time for Printing on Printer 1 = Number of Pages on Printer 1 / Printing Rate of Printer 1
Number of Pages on Printer 2 = Total Number of Pages - Number of Pages on Printer 1
Time for Printing on Printer 2 = Number of Pages on Printer 2 / Printing Rate of Printer 2
Total Printing Time = Time for Printing on Printer 1 + Time for Printing on Printer 2
Time Print Job Finished = Starting Print Time + Total Printing Time

Solution Steps:

time P1 = $\frac{305 p}{15 p/min}$ = 20.33 min = 1220 sec.

Pages P2 = 720 - 305 = 415 p.

time P2 = $\frac{415 p}{50 p/min}$ = 8.3 min = 498 sec

total P time = 1220 sec + 498 sec = 1718 sec

1718 sec = 28.63 min = 28 min 40 sec

Time finished = 10 min 15 sec + 28 min 40 sec
= 39 min 55 sec past the hour

Calculations (Show results in "Solutions Steps"):

305 | 15
050 | 20.33
50
60
0000
12198X
121980

720
305
415 | 50
15 | 83
60
664
498
1220
1718

1718 | 60
51 | 28.63
38
20
20

ANSWER: 39 min 55 sec past the hour

(* S261P12 B12 *)

GIVENS:
UNKNOWNS:
305 pages / 15 pages per min = 20.33 min
20.33 min = 1220 sec
720 - 305 = 415 pages
415 pages / 50 pages per min = 8.3 min
8.3 min = 498 sec
1220 sec + 498 sec = 1718 sec
1718 sec = 28.63 min
28.63 min = 28 min 40 sec
10 min 15 sec + 28 min 40 sec = 39 min 5 sec
ANSWER = 39 min 5 sec

PLEASE PRINT YOUR SOLUTION CLEARLY. Name: [redacted]

Problem 18. On a 600-mile motor trip, Bill averaged 45 miles per hour for the first 285 miles and 50 miles per hour for the remainder of the trip. If he started at 7:00 a.m., at what time did he finish the trip (to the nearest minute)?

Solution Steps:

Time 1st half = 45 | 285
= 6.33 hours
Dist. 2nd half = 600 − 285
= 315 miles
Time 2nd half = 50 | 315
= 6.3 hours
Total time for trip = 12.63 hours ↓ convert to min
+ 7:00 am
12:40
19:40

Calculations (Show results in "Solutions Steps"):

285
45 mph | 50 mph
600

6.33
45 | 285.0
270
150
135
15

~~800~~ 900
− 285
315

6.3
50 | 315.0

2 | 63 | 150 00
9
567
330

ANSWER: 19:40 or 7:40 pm

(* S067P01 A18 *)

GIVENS:
UNKNOWNS:
time_first_half = 285 / 45 mph
285 / 45 mph = 6.33 hr
distance_second_half = 600 - 285
600 - 285 = 315 miles
time_second_half = 315 / 50
315 / 50 mph = 6.3 hr
Total_Time = 12.63 hr
12.63 hr = 12 hr 40 min
7:00 am + 12:40 = 19:40
ANSWER = 19:40

Appendix C

Scoring Rubric and Keys

GRE Quantitative Constructed-Response Scoring Rubric

1. If the student provides two or more solutions, consider only the best one. In general, do not deduct credit if the student explicitly corrects errors.

2. Consider all available information including that in the "Calculations Space."

3. If only the final answer is present and it is correct, give full credit because there is no process on which to make any other decision. In all other cases, the total score for the problem is the sum of the scores for each goal.

4. Each goal is worth 3 points. Deduct points as follows:

a. Deduct 3 points if the goal is missing and is not implicitly satisfied. A goal is considered missing when there is no reasonable attempt to solve for it. A goal is considered to be implicitly satisfied if it can be inferred from other parts of the solution.

b. Deduct 2 points if the goal is present but contains an uncorrected structural error (e.g., inverting the dividend and the divisor, confusing operators). For a goal to be considered present but structurally incorrect, it must be clearly evident that the student is making an attempt--however misguided--to solve the goal (thereby showing awareness that solving for that goal is a step in the problem's solution process). The minimal evidence needed to indicate such an attempt is the presence of a reasonable expression bound to a label that can be unambiguously associated with that goal.

c. Deduct 1 point for each computational error within a present goal. Count as computational errors miscalculations (including those beyond the required level of precision), transcription errors (values incorrectly copied from one part of the problem to another), errors in copying a given from the problem statement, conversion errors (unless otherwise indicated), and, for the last goal only, failing to reduce the final answer to a single value. Only deduct for the same computational error once. For all computational errors, carry through the result to subsequent goals, giving full credit to those subsequent goals if they are structurally and computationally correct given their incorrect input.

d. Deduct 1 point for failing to carry the result of a goal to the required level of precision (i.e., two decimal places or the precision required by the individual problem, whichever is greater).

e. Deduct 0 points if the goal is present and correct. A goal should be considered to be present and correct if (1) the result and the method are correct, (2) the result is correct and the method is not identifiably faulty, or (3) the method is correct and the result is incorrect only because the inputs to the goal appropriately came from a previous goal that incorrectly computed those inputs.

In making the above deductions, try to distinguish between errors that can be explained by a single fault and those that are composites of two or more

faults. The following example could be conceived as a single error in which the student has mistakenly converted a decimal representation to time. This would constitute a single error for which 1 point would be deducted.

Time1 = 10.67
Time1 = 11 hr 7 min

In contrast, the following production could be interpreted as two separable errors, one in failing to round 10.66 to 10.67 (the result of 800/75), and the second in confusing decimal and time representations. For this goal, one point would be deducted for each of these computational mistakes.

Time1 = 800/75
Time1 = 11 hr 6 min

5. Unless the final answer (the value on the ANSWER line) is redundant with the culminating value in the student's solution, treat this final answer as part of the solution proper. That is, in many student solutions the ANSWER line value is <u>not</u> redundant but instead represents the result of the student's last goal. Such values should be included in scoring that goal.

6. Treat as equivalent the various operational notations (e.g., *, x, (), ·); mixed numbers and improper fractions (e.g., $8^1/_3$ and $^{25}/_3$); numbers with and without units (400 and 400 doses); and percentages, decimals, and fraction equivalents (e.g., $^1/_4$%, .25%, .0025, and $^1/_{400}$).

7. Treat as correct a goal that is satisfied except for the presence of a unit conversion if that conversion is made in a subsequent goal. In the example below, treat equivalently the conversion of hours to hours and minutes whether it occurs in goal #5, goal #4, or in goals #1 and #2.

Problem: On a 600-hundred mile motor trip, Bill averaged 45 miles per hour for the first 285 miles and 50 miles per hour for the remainder of the trip. If he started at 7:00 a.m., at what time did he finish the trip (to the nearest minute)?

a. Time 1 = 285 miles / 45 miles per hour
   Time 1 = 6.33 hours (6.33 hours = 6 hours and 20 minutes)
b. Distance 2 = 600 miles - 285 miles
   Distance 2 = 315 miles
c. Time 2 = 315 miles / 50 mile per hour
   Time 2 = 6.3 hours (6.3 hours = 6 hours and 18 minutes)
d. Total time = 6.33 hours + 6.3 hours
   Total time = 6 hours 20 min + 6 hours 18 min
   Total time = 12 hours 38 min
e. End time = 7:00 am + 12 hours 38 min (7:00 am + 12.63 hrs = 7:38 pm)
   End time = 7:38 pm

8. In some cases, the scoring key for a problem presents two alternative goal decompositions. Score the examinee response according to the decomposition that best characterizes the response. Be sure to use the same maximum scores and the same point deduction rules <u>regardless</u> of the decomposition being used to score the response. Under this rule, partially correct solutions that

follow more efficient decompositions will generally receive more points than similar quality solutions following less efficient decompositions.

9. The minimum score for a goal is 0 as is the minimum total score for a solution.

Scoring Keys

Two-Goal Problems

How many minutes will it take to fill a 2,000-cubic-centimeter tank if water flows in at the rate of 20 cubic centimeters per minute and is pumped out at the rate of 4 cubic centimeters per minute?

ANSWER = 125 minutes

Decomposition #1

1. Net filling rate = 20 cc per minute - 4 cc per minute
   Net filling rate = 16 cc per minute
2. Time to fill tank = 2,000 cc/16 cc per minute
   Time to fill tank = 125 min

Maximum possible score = 6

Decomposition #2

1. Time to fill alone = 2000 cc/20 cc p min
   Time to fill along = 100 minutes
   Time to empty alone = 2000 cc/4 cc p min
   Time to empty alone = 500 minutes
2. Net fill time * (1 tank filling/100 min - 1 tank filling/500 min) = 1 tank filling
   Net fill time * (4 tank fillings/500 min) = 1 tank filling
   Net fill time = 500/4 minutes
   Net fill time = 125 minutes

Maximum possible score = 6

One of two outlets of a small business is losing $500 per month while the other is making a profit of $1750 per month. In how many months will the net profit of the small business be $35,000?

ANSWER = 28 months

Decomposition #1

1. Net monthly profit = $1750 per month - $500 per month
   Net monthly profit = $1250 per month
2. Months to reach target profit = $35,000/$1250 per month
   Months to reach target profit = 28 months

Maximum possible score = 6

Decomposition #2

1. Time for outlet 1 loss alone = $35,000/$500 per month
   Time for outlet 1 loss alone = 70 months
   Time for outlet 2 profit alone = $35,000/$1750 per month
   Time for outlet 2 profit alone = 20 months
2. Time for net profit * (1 unit target profit/20 months - 1 unit target profit/70 months)
   = 1 unit target profit
   Time for net profit * (50 units target profit/1400 months) = 1 unit target profit
   Time for net profit = 1400/50 months
   Time for net profit = 28 months

Maximum possible score = 6

A specialty chemical company has patented a chemical process that involves 2 reactions. Reaction 1 generates 24 grams of molecule B per minute and reaction 2 consumes 5 grams of molecule B per minute. If 4,560 grams of molecule B are desired as a product of this process, how many minutes must it continue?

ANSWER = 240 minutes

Decomposition #1

1. Net amount of B per minute = 24 gm per min - 5 gm per min
   Net amount of B per minute = 19 gm per min
2. Time for desired amount = 4560 gm/19 gm per min
   Time for desired amount = 240 min

Maximum possible score = 6

Decomposition #2

1. Time for unit product gain from reaction 1 alone = 4560 gm/24 gm p min
   Time for unit product gain from reaction 1 alone = 190 minutes
   Time for unit product loss from reaction 2 alone = 4560 gm/5 gm p min
   Time for unit product loss from reaction 2 alone = 912 minutes
2. Time for net product * (1 unit product/190 min - 1 unit product/912 min)
   = 1 unit product
   Time for net product * (722 units product/173,280 min) = 1 unit product
   Time for net product = 173,280/722 minutes
   Time for net product = 240 minutes

Maximum possible score = 6

$3.50 in tolls is received each minute at an automated toll booth while the rate at a booth with an operator is $2.80 each minute. How many minutes elapse before the automated booth receives $14.00 more in tolls than does the person-operated booth?

ANSWER = 20 minutes

Decomposition #1

1. Tolls per minute = $3.50 per minute - $2.80 per minute
   Tolls per minute = $.70 per minute
2. Time for $14 lead = $14/$.70 per minute
   Time for $14 lead = 20 minutes

Maximum possible score = 6

Decomposition #2

1. Time for $14 from auto booth alone = $14/$3.50 per min
   Time for $14 from auto booth alone = 4 minutes
   Time for $14 from manual booth alone = $14/$2.80 per minute
   Time for $14 from manual booth alone = 5 minutes
2. Time for $14 lead * (1 unit of $14/4 minutes - 1 unit of $14/5minutes) = 1 unit of $14
   Time for $14 lead * (1 unit of $14/20 minutes) = 1 unit of $14
   Time for $14 lead = 20/1 minutes
   Time for $14 lead = 20 minutes

Maximum possible score = 6

Three-Goal Problems

Money in a certain investment fund earns an annual dividend of 5 percent of the original investment. In how many years will an initial investment of $750 earn total dividends equal to the original investment?

ANSWER = 20 years

Decomposition #1

1. 5% = .05
2. Yearly earning = .05 * $750
   Yearly earning = $37.50
3. Time to earn investment = $750/$37.50 per year
   Time to earn investment = 20 years

Maximum possible score = 9

Decomposition #2

1. 5% dividend * X years = 100% dividend
2. X years = 100% dividend/5% dividend
   X = 20 years

Maximum possible score = 9

On every $150 load of cement it delivers to a construction site, Acme Cement Company earns a 4 percent profit. How many loads must it deliver to the site to earn $150 in profit?

ANSWER = 25 loads

Decomposition #1

1. 4% = .04
2. Profit per load = .04 * $150
   Profit per load = $6
3. Loads for target profit = $150/$6 per load
   Loads for target profit = 25 loads

Maximum possible score = 9

Decomposition #2

1. 4% profit * X loads = 100% profit
2. X loads = 100% profit/4% profit
   X = 25 loads

Maximum possible score = 9

A graphics designer earns 2% of a $1500 yearly bonus for each shift of overtime she works. How many shifts of overtime must she work to earn the equivalent of the entire yearly bonus?

ANSWER = 50 shifts

Decomposition #1

1. 2% = .02
2. Amount earned per shift = .02 * $1500
   Amount earned per shift = $30
3. Number of shifts for bonus = $1500/$30 per shift
   Number of shifts for bonus = 50 shifts

Maximum possible score = 9

Decomposition #2

1. 2% bonus * X shifts = 100% bonus
2. X shifts = 100% bonus/2% bonus
   X = 50 shifts

Maximum possible score = 9

The active ingredient is 0.25 percent of a 3-ounce dose of a certain cold remedy. What is the number of doses a patient must take before receiving the full 3 ounces of the active ingredient?

ANSWER= 400 doses

Decomposition #1

1. 0.25% = .0025
2. Active Ingredient per dose = .0025 * 3 oz
   Active Ingredient per dose = .0075 oz
3. Number of doses required = 3 oz/.0075 oz per dose
   Number of doses required = 400 doses

Maximum possible score = 9

Decomposition #2

1. .25% * X doses = 100% dose
2. X dose = 100% dose/.25% dose
   X = 400 doses

Maximum possible score = 9

Five-Goal Problems

On a 600-hundred mile motor trip, Bill averaged 45 miles per hour for the first 285 miles and 50 miles per hour for the remainder of the trip. If he started at 7:00 a.m., at what time did he finish the trip (to the nearest minute)?

ANSWER = 7:38 pm

1. Time 1 = 285 miles / 45 miles per hour
   Time 1 = 6.33 hours
2. Distance 2 = 600 miles - 285 miles
   Distance 2 = 315 miles
3. Time 2 = 315 miles / 50 mile per hour
   Time 2 = 6.3 hours
4. Total time = 6.33 hours + 6.3 hours
   Total time = 6 hours 20 min + 6 hours 18 min
   Total time = 12 hours 38 min
5. End time = 7:00 am + 12 hours 38 min
   End time = 7:38 pm

Maximum possible score = 15

800 gallons of a 2,400 gallon tank flow in at the rate of 75 gallons per hour through a clogged hose. After the hose is unclogged, the rest of the tank is filled at the rate of 250 gallons per hour. At what time to the nearest minute will the filling of the tank be finished if it starts at 5:30 a.m.?

ANSWER = 10:34 pm

1. Filling time 1 = 800 gal/75 gal per hour
   Filling time 1 = 10 and 2/3 hrs
2. Filling amount 2 = 2400 gal - 800 gal
   Filling amount 2 = 1600 gal
3. Filling time 2 = 1600 gal / 250 gal per hour
   Filling time 2 = 6 and 4/10 hrs
4. Total filling time = 10 and 2/3 hrs + 6 and 4/10 hrs
   Total filling time = 10 hrs 40 min + 6 hrs 24 min
   Total filling time = 17 hrs 4 min
5. Ending time for filling = 5:30 am + 17 hrs 4 min
   Ending time for filling = 10:34 pm

Maximum possible score = 15

Of the 720 pages of printed output of a certain program, 305 pages are printed on a printer that prints 15 pages per minute and the rest are printed on a printer that prints at 50 pages per minute. If the printers run one after the other and printing starts at 10 minutes and 15 seconds after the hour, at what time to the nearest second after the hour will the printing be finished?

ANSWER: 38 minutes 53 seconds

1. Time on printer 1 = 305 pages/15 pages per minute
   Time on printer 1 = 20 and 1/3 minutes
2. Pages on printer 2 = 720 pages - 305 pages
   Pages on printer 2 = 415 pages
3. Time on printer 2 = 415 pages/50 pages per minute
   Time on printer 2 = 8.3 min
4. Total printing time = 20 and 1/3 min + 8.3 min
   Total printing time = 20 min 20 sec + 8 min 18 sec
   Total printing time = 28 min 38 sec
5. Time finished = 10 min 15 sec + 28 min 38 sec
   Time finished = 38 min 53 sec

Maximum possible score = 15

A Department of Transportation road crew paves the 15 mile city portion of a 37.4 mile route at the rate of 1.8 miles per day and paves the rest of the route, which is outside the city, at a rate of 2.1 miles per day. If the Department of Transportation starts the project on day 11 of its work calendar, on what day of its work calendar will the project be completed?

ANSWER= 30th day or 29th day (to recognize that the work might have begun on the morning of the 11th day)

1. Time for portion 1 = 15 miles/1.8 miles per day
   Time for portion 1 = 8 and 1/3 days
2. Portion 2 distance = 37.4 miles - 15 miles
   Portion 2 distance = 22.4 miles
3. Time for portion 2 = 22.4 miles/2.1 miles per day
   Time for portion 2 = 10 and 2/3 days
4. Total time = 8 and 1/3 days + 10 and 2/3 days
   Total time = 19 days
5. Completion day = 11th day + 19 days
   Completion day = 30th day or 29th day)

Maximum possible score = 15